

# Causal Inference on Multivariate Mixed Type Data

ALEXANDER MARX, Max Planck Institute for Informatics and Saarland University, Saarbrücken, Germany

JILLES VREEKEN, Max Planck Institute for Informatics and Saarland University, Saarbrücken, Germany

Given data over the joint distribution of two univariate or multivariate random variables  $X$  and  $Y$  of mixed or single type data, we consider the problem of inferring the most likely causal direction between  $X$  and  $Y$ . We take an information theoretic approach, from which it follows that first describing the data over *cause* and then that of *effect* given *cause* is shorter than the reverse direction.

For practical inference, we propose a score for causal models for mixed type data based on the Minimum Description Length (MDL) principle. In particular, we model dependencies between  $X$  and  $Y$  using classification and regression trees. Inferring the optimal model is NP-hard, and hence we propose CRACK, a fast greedy algorithm to infer the most likely causal direction directly from the data.

Empirical evaluation on synthetic, benchmark, and real world data shows that CRACK reliably and with high accuracy infers the correct causal direction on both univariate and multivariate cause–effect pairs over both single and mixed type data.

CCS Concepts: •Information systems → Data mining; •Mathematics of computing → Information theory; Causal networks;

Additional Key Words and Phrases: MDL, causal inference, decision trees, mixed-type data

## ACM Reference format:

Alexander Marx and Jilles Vreeken. 2017. Causal Inference on Multivariate Mixed Type Data. 1, 1, Article 1 (January 2017), 16 pages. DOI: 10.475/123.4

## 1 INTRODUCTION

Telling cause from effect is one of the core problems in science. It is often difficult, expensive, or impossible to obtain data through randomized trials, and hence we often have to infer causality from, what is called, observational data [20]. We consider the setting where, given data over the joint distribution of two random variables  $X$  and  $Y$ , assuming no hidden confounders, we have to infer the most likely causal direction between  $X$  and  $Y$ . In other words, our task is to identify whether it is more likely that  $X$  causes  $Y$ , or vice versa, that  $Y$  causes  $X$ , or that the two are merely correlated.

In practice,  $X$  and  $Y$  do not have to be of the same type. The altitude of a location (real-valued), for example, determines whether it is a good habitat (binary) for a mountain hare. In fact, whether a location is a good habitat or not for an animal is not caused by a single aspect, but by a *combination* of conditions. We are therefore interested in the general case where  $X$  and  $Y$  may be of any cardinality, i.e. univariate or multivariate, and may be single or mixed-type. That is, both  $X$  and  $Y$  may consist of a mix of binary, categorical, discrete numeric, or continuous real-valued attributes.

To the best of our knowledge there exists no method for this general setting. Causal inference based on conditional independence tests, for example, requires three variables, and cannot decide between  $X \rightarrow Y$  and  $Y \rightarrow X$  [20]. All

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

© 2017 Copyright held by the owner/author(s). Manuscript submitted to ACM

existing methods that consider two variables are only defined for single-type pairs. Additive Noise Models (ANMs), for example, have only been proposed for univariate pairs of real-valued [23] or discrete variables [22], and similarly so for methods based on the independence of  $P(X)$  and  $P(Y | X)$  [16, 27]. Trace-based methods require both  $X$  and  $Y$  to be strictly multivariate real-valued [2, 9], and whereas ERGO [32] also works for univariate pairs, these have to be real-valued. We refer the reader to Sec. 6 for a more detailed overview of related work.

Our approach is based on algorithmic information theory. That is, we follow the postulate that if  $X \rightarrow Y$ , it will be easier—in terms of Kolmogorov complexity—to first describe the data of  $X$ , and then describe the data of  $Y$  given  $X$ , than vice-versa [1, 11, 32]. In other words: causal inference by compression. Kolmogorov complexity is not computable, but can be approximated through the Minimum Description Length (MDL) principle [5, 25], by which we can instantiate this framework in practice [1].

To this end, we define an MDL score for coding forests, a model class where a model consists of classification and regression trees. By allowing dependencies from  $X$  to  $Y$ , or vice versa, we can measure the difference in complexity between  $X \rightarrow Y$  and  $Y \rightarrow X$ . Discovering a single optimal decision tree is already NP-hard [18], and hence we cannot efficiently discover the coding forest that describes the data most succinctly. We therefore propose CRACK, an efficient greedy algorithm for discovering good models from data.

Through extensive empirical evaluation on synthetic, benchmark, and real-world data we show that CRACK performs very well in practice. It infers the correct causal direction with high accuracy, even for weak dependencies. It performs at least as well as existing methods for univariate single-type pairs, and outperforms the state of the art on multivariate pairs. It is also very fast, taking less than 4 seconds over any pair in our experiments.

The main contributions of this paper are as follows.

- (a) we propose the first framework for causal inference on univariate and multivariate mixed-type data,
- (b) define an MDL score for the model class of coding trees,
- (c) give the efficient CRACK algorithm,
- (d) provide extensive experimental results, and
- (e) make our implementation and all used data available.

The paper is structured as usual. We introduce our notation in Sec. 2, and give a brief primer to causal inference by Kolmogorov complexity and the Minimum Description Length principle in Sec. 3. We formalize our MDL score in Sec. 4, and present the efficient CRACK algorithm for finding good models in Sec. 5. Related work is discussed in Sec. 6, and we evaluate CRACK empirically in Sec. 7. We round up with discussion in Sec. 8 and conclude in Sec. 9.

## 2 NOTATION

In this work we consider data  $D$  over the joint distribution of random variables  $X$  and  $Y$ . The database  $D$  contains  $n$  records and a set of  $A$  of  $|A| = |X| + |Y| = m$  attributes,  $a_1, \dots, a_m \in A$ . An attribute  $a$  has a type  $type(a)$  where  $type(a) \in \{binary, categorical, numeric\}$ . We will often refer to binary and categorical attributes as *nominal* attributes. The size of the domain of an attribute  $a$  is defined as

$$|dom(a)| = \begin{cases} \#values & \text{if } type(a) \text{ is nominal} \\ \frac{\max(a) - \min(a)}{res(a)} + 1 & \text{if } type(a) \text{ is numeric,} \end{cases}$$

where  $res(a) \in (0, 1]$  is the resolution at which the data over attribute  $a$  was recorded. For example, a resolution of 1 means that we consider integers, and a resolution of 0.01 means the data of  $a$  was recorded with a precision of up to a hundredth.

We will consider decision and regression trees. In general, a tree  $T$  consist of  $|T|$  nodes. We identify internal nodes as  $v \in int(T)$ , and leaf nodes as  $l \in lvs(T)$ . A leaf node  $l$  contains  $|l|$  data points.

All logarithms are to base 2, and by convention we say  $0 \log 0 = 0$ .

### 3 CAUSAL INFERENCE BY COMPRESSION

In this paper we pursue the goal of causal inference by compression. Below we give a short introduction to the key concepts we use.

#### 3.1 Kolmogorov Complexity, a brief primer

The Kolmogorov complexity of a finite binary string  $x$  is the length of the shortest binary program  $p^*$  for a Universal Turing machine  $\mathcal{U}$  that generates  $x$ , and then halts [13, 15]. Formally, we have

$$K(x) = \min\{|p| \mid p \in \{0, 1\}^*, \mathcal{U}(p) = x\} .$$

Simply put,  $p^*$  is the most succinct *algorithmic* description of  $x$ , and the Kolmogorov complexity of  $x$  is the length of its ultimate lossless compression. Conditional Kolmogorov complexity,  $K(x \mid y) \leq K(x)$ , is then the length of the shortest binary program  $p^*$  that generates  $x$ , and halts, given  $y$  as input. For more details see [15].

#### 3.2 Causal Inference by Complexity

The problem we consider is to infer, given data over two correlated variables  $X$  and  $Y$ , whether  $X$  caused  $Y$ , whether  $Y$  caused  $X$ , or whether  $X$  and  $Y$  are only correlated. As is common, we assume causal sufficiency. That is, we assume there exists no hidden confounding variable  $Z$  that is the common cause of both  $X$  and  $Y$ .

The Algorithmic Markov condition, as recently postulated by Janzing and Schölkopf [11], states that factorizing the joint distribution over *cause* and *effect* into  $P(\text{cause})$  and  $P(\text{effect} \mid \text{cause})$ , will lead to simpler—in terms of Kolmogorov complexity—models than factorizing it into  $P(\text{effect})$  and  $P(\text{cause} \mid \text{effect})$ . Formally, they say that if  $X$  causes  $Y$ ,

$$K(P(X)) + K(P(Y \mid X)) < K(P(Y)) + K(P(X \mid Y)) .$$

This *model-driven* postulate reasons about the complexity of given true distributions  $P(\cdot)$ . In practice we do not have access to these distributions, however, and only to empirical data.

Budhathoki & Vreeken [1] showed that we can define causality in terms of Kolmogorov complexity over the observed data. Loosely speaking, this postulate says that it will be simpler to first *describe* the *data* over *cause*, and then describe the *data* over *effect* given the data over *cause*, than vice versa. That is, we do not reason about complexities of distributions alone, but rather on the complexities  $K(X)$  and  $K(Y \mid X)$  of observed data  $X$  and  $Y$  over  $X$  and  $Y$ .

Vreeken [32] proposed to consider the relative *conditional* complexity of the data over  $X$  and  $Y$  as causal indicator  $\delta_{X \rightarrow Y}$ , with

$$\delta_{X \rightarrow Y} = \frac{K(Y \mid X)}{K(Y)} , \tag{1}$$

where we normalize by  $K(Y)$  to avoid bias towards simple objects, i.e. those with low  $K(X)$  or  $K(Y)$ . Intuitively the score corresponds to the remaining complexity of  $Y$  knowing  $X$ . It will be 1 when  $X$  contains no information towards  $Y$ , i.e. when  $X$  is *algorithmically independent* of  $Y$ , and will be close to 0 if  $X$  contains all information of  $Y$ . We infer that  $X$  is a likely algorithmic cause for  $Y$ , denoted by  $X \rightarrow Y$ . Alternatively, if  $\delta_{Y \rightarrow X} < \delta_{X \rightarrow Y}$  we infer  $Y \rightarrow X$  as the most likely direction.

Budhathoki & Vreeken [1] show that  $\delta_{X \rightarrow Y}$  has merit, yet is biased towards the more *complex* object. To alleviate, they propose to consider the relative joint complexity,

$$\Delta_{X \rightarrow Y} = \frac{K(X) + K(Y | X)}{K(X) + K(Y)}, \quad (2)$$

While in general the symmetry of information,  $K(x) + K(y | x) = K(y) + K(x | y)$ , holds up to an additive constant [15], Janzing and Schölkopf [11] showed it does *not* hold when  $X$  causes  $Y$ , or vice versa. This asymmetry allows us to infer that  $X \rightarrow Y$  as the most likely causal direction if  $\Delta_{X \rightarrow Y} < \Delta_{Y \rightarrow X}$ , and vice versa.

Due to the halting problem, Kolmogorov complexity is not computable. We can approximate it, however, via the Minimal Description Length (MDL) principle [5, 15].

### 3.3 MDL, a brief primer

The Minimum Description Length (MDL) principle [5, 25] is a practical variant of Kolmogorov Complexity. Intuitively, instead of all programs, it considers only those programs that we know that output  $x$  and halt. Formally, given a model class  $\mathcal{M}$ , MDL identifies the best model  $M \in \mathcal{M}$  for data  $D$  as the one minimizing

$$L(D, M) = L(M) + L(D | M),$$

where  $L(M)$  is the length in bits of the description of  $M$ , and  $L(D | M)$  is the length in bits of the description of data  $D$  given  $M$ . This is known as two-part MDL. There also exists one-part, or *refined* MDL, where we encode data and model together. Refined MDL is superior in that it avoids arbitrary choices in the description language  $L$ , but is computable only for certain model classes. Note that in either case we are only concerned with code *lengths* — our goal is to measure the *complexity* of a dataset under a model class, not to actually compress it [5].

### 3.4 Causal Inference by MDL

For causal inference by MDL, we will need to approximate both  $K(X)$  and  $K(Y | X)$ . For the former, we need to consider the class  $\mathcal{M}_X$  of models  $M_X$  that describe data  $X$  without knowledge of  $Y$ , while for the latter we need to consider class  $\mathcal{M}_{Y|X}$  of models  $M_{Y|X}$  that describe the data of  $Y$  knowing the data of  $X$ .

That is, we are after the *causal* model  $M_{X \rightarrow Y} = (M_X, M_{Y|X})$  from the class  $\mathcal{M}_{X \rightarrow Y} = \mathcal{M}_X \times \mathcal{M}_{Y|X}$  that best describes the data over  $X$  and  $Y$ . By MDL, we identify the optimal model  $M_{X \rightarrow Y} \in \mathcal{M}_{X \rightarrow Y}$  for data  $D$  over  $X$  and  $Y$  as the one minimizing

$$L(D, M_{X \rightarrow Y}) = L(X, M_X) + L(Y, M_{Y|X} | X),$$

where the encoded length of  $X$  and model  $M_X$  is defined as

$$L(X, M_X) = L(M_X) + L(X | M_X),$$

and we define accordingly

$$L(Y, M_{Y|X} | X) = L(M_{Y|X}) + L(Y | M_{Y|X}, X)$$

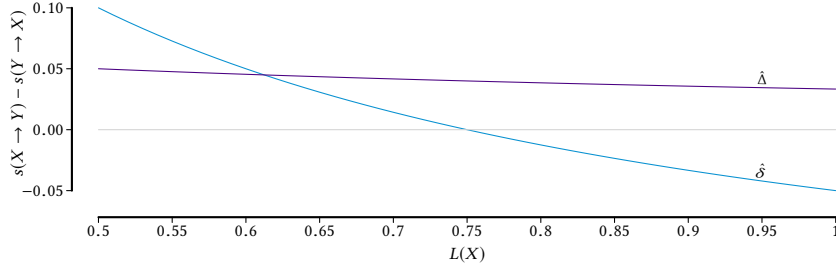


Fig. 1. Effect of the domain size of  $X$  on  $\hat{\Delta}$  and  $\hat{\delta}$  by keeping  $L(Y)$  and the gains in compression constant. If  $s(X \rightarrow Y) - s(Y \rightarrow X)$  of the corresponding score is  $< 0$  it infers  $X \rightarrow Y$  and for  $> 0$  it infers  $Y \rightarrow X$ .

as the encoded length of data  $Y$  and model  $M_{Y|X}$  given data  $X$ .

To identify the most likely causal direction between  $X$  and  $Y$  by MDL we can now simply rewrite Eq. (3.2) and Eq. (3.2) as

$$\hat{\delta}_{X \rightarrow Y} = \frac{L(Y, M_{Y|X} | X)}{L(Y, M_Y)}, \text{ and}$$

$$\hat{\Delta}_{X \rightarrow Y} = \frac{L(X, M_X) + L(Y, M_{Y|X} | X)}{L(X, M_X) + L(Y, M_Y)}.$$

Similar to the original scores, we infer that  $X$  is a likely cause of  $Y$  if  $\hat{\delta}_{X \rightarrow Y} < \hat{\delta}_{Y \rightarrow X}$ , and vice versa, and analogue for and the relative joint complexities  $\hat{\Delta}_{X \rightarrow Y}$  and  $\hat{\Delta}_{Y \rightarrow X}$ .

To use these causal indicators in practice, we need to define a casual model class  $\mathcal{M}_{X \rightarrow Y}$ , how to encode a model  $M \in \mathcal{M}$  in bits, and how to encode a dataset  $D$  using a model  $M$ . This we will do in Section 4. First, we discuss the merits of both scores.

### 3.5 Robustness of the $\hat{\delta}$ and $\hat{\Delta}$

Vreeken & Budhathoki [1] observed that  $\hat{\delta}$  is biased towards objects of higher complexity, and show that the relative joint complexity score,  $\hat{\Delta}$ , performs better for binary  $X$  and  $Y$  with asymmetric cardinality. We observe that the complexity of the distribution of a continuous real-valued attributes also plays a role, and that  $\hat{\Delta}$  is close to invariant to this.

To illustrate, let us consider the following example. Suppose  $L(X) = L(Y) = 0.5$ , and let the gain in compression for  $X \rightarrow Y$  be  $L(Y) - L(Y | X) = 0.1$ , and the gain in compression for  $Y \rightarrow X$  be 0.15. Since,  $X$  can be expressed most succinctly as an effect of  $Y$ , both scores will infer  $Y \rightarrow X$ . Next, let us consider a different setting, where we adjust the complexity of  $X$ , for example by adjusting the resolution, such that  $L(X) = 1.0$ , but keeping all other complexities the same. This means relative to the initial size of  $X$  the gain in compression  $X \rightarrow Y$  drops from 0.3 to 0.15. In contrast, the relative gain for  $Y \rightarrow X$  is 0.2 in both examples. Whereas  $\hat{\delta}$  now infers  $X \rightarrow Y$ ,  $\hat{\Delta}$  still infers  $Y \rightarrow X$ . We plot this relation in Figure 1, using the values as discussed and varying  $L(X)$  from 0.5 to 1. The y-axis shows the difference between the scores.

In sum, both scores have merit.  $\hat{\Delta}$  is more robust to asymmetries in the cardinality of  $X$  and  $Y$ , and  $\hat{\delta}$  is more robust to unbalanced domain sizes of symmetric, e.g. univariate, real-valued  $X$  and  $Y$ .

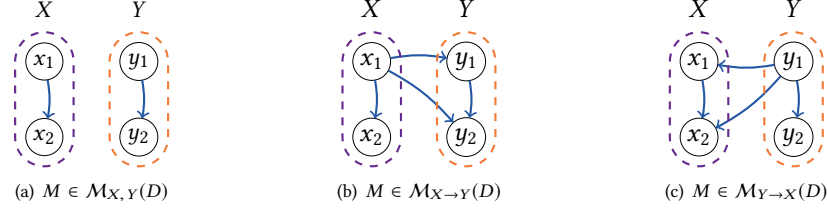


Fig. 2. Toy examples of valid models  $M$  for different model classes  $\mathcal{M}$ . An edge from a node  $u$  to a node  $v$  means that  $v$  depends on  $u$ .

#### 4 MDL FOR TREE MODELS

As models we consider tree models, or, *coding forests*. A coding forest  $M$  contains per attribute  $a_i \in A$  one coding tree  $T_i$ . A coding tree  $T_i$  is simply a binary tree that encodes the values of  $a_i$  in its leaves, splitting or regressing the data of  $a_i$  on attribute  $a_j$  ( $i \neq j$ ) in its internal nodes to encode the data of  $a_i$  more succinctly.

We encode the data over every attribute  $a_i$  with its corresponding coding tree  $T_i$ . The total encoded length of data  $D$  and  $M$  then is

$$L(D, M) = \sum_{a_i \in A} L(T_i),$$

which corresponds to the sum of costs of the individual trees.

To ensure lossless decoding, there needs to exist an order on the trees  $T \in M$  such that we can transmit these one by one. In other words, in a *valid* tree model there are no cyclic dependencies between the trees  $T \in M$ , and a valid model can hence be represented by a DAG. Let  $\mathcal{M}(D)$  be the set of all valid tree models for  $D$ , that is,  $M \in \mathcal{M}(D)$  is a set of  $|A|$  trees such that the data types of the leafs in  $T_i$  corresponds to the data type of attribute  $a_i$ , and its dependency graph is acyclic.

We write  $\mathcal{M}_X(D)$  to denote the subset of valid coding forests for  $D$  where we only allow dependencies between attributes in  $X \subseteq A$ , that is  $\mathcal{M}_X(D) \subseteq \mathcal{M}_A(D) = \mathcal{M}(D)$ . Similar, we identify by  $\mathcal{M}_{X,Y}(D)$  that subset of valid coding forests for  $D$  where we allow dependencies between attributes  $X$ , and between attributes  $Y$ , but not in between. An example model  $M \in \mathcal{M}_{X,Y}(D)$  is plotted in Figure 2 (a). Third, we define  $\mathcal{M}_{Y|X}(D)$  as the set of valid models for  $D$  where we allow attributes of  $Y$  to a-cyclically depend on each other, as well as on attributes from  $X$ . Last, but not least, as we are concerned with causal models, we write  $\mathcal{M}_{X \rightarrow Y}(D)$  for the set of all valid models for  $D$  where we allow attributes of  $X$  to depend on each other, and attributes from  $Y$  to depend on either  $X$  or  $Y$ . Example models that are valid for  $\mathcal{M}_{X \rightarrow Y}(D)$  or for its reverse  $\mathcal{M}_{Y \rightarrow X}(D)$  are given in Figures 2 (b) and (c).

*Cost of a Tree.* The encoded cost of a tree consists of two parts. First, we transmit the topology of the tree, and then how the data is separated or transformed. Second, we transmit the data in the leaves of the tree. Formally, we have

$$L(T) = |T| + \sum_{v \in \text{int}(T)} (1 + L(v)) + \sum_{l \in \text{ls}(T)} L(l),$$

where per node we need one bit to indicate if it is an internal or a leaf node. An internal node can either split the data, or apply regression. We identify the type of the node with one bit.

*Cost of Splitting.* The encoded length of a split node  $v$  is

$$L_{split}(v) = \log |A| + \begin{cases} \log |dom(a_j)| & \text{if } a_i \text{ is categorical} \\ \log |dom(a_j) - 1| & \text{else,} \end{cases}$$

whereas we first identify in  $\log |A|$  which attribute  $a_j$  the node splits the data of  $a_i$  on, and second the condition  $a_j = x$  on which we split the data.

For categorical data, we identify the attribute value on which we split without any preference. Hence, the costs are  $\log |dom(a_j)|$ . For numeric attributes we need to identify the cut point on the candidate. A cut point lies between two consecutive values in the domain of the candidate. As we do not have any preference between which values the split is set, we encode the costs accordingly using  $\log |dom(a_j) - 1|$  bits. For binary attributes it is invariant whether we split on  $x = 1$ , or  $x = 0$ , and hence the cost is  $\log |dom(a_j) - 1| = \log |2 - 1| = 0$ .

*Cost of Regressing.* For a regression node we also first encode the target attribute, and then the parameters of the regression, i.e.

$$L_{reg}(v) = \log |A| + \sum_{\phi \in \Phi(v)} 1 + L_{\mathbb{N}} \left( \left\lceil \frac{|\phi|}{res(a_i)} \right\rceil + 1 \right),$$

where  $\Phi(v)$  denotes the set of parameters for the regression. For linear regression, it consists of  $\alpha$  and  $\beta$ , while for quadratic regression it further contains  $\gamma$ . To describe each parameter  $\phi \in \Phi$  we first encode its sign using one bit, and then encode its absolute value in the resolution of  $a_i$  using  $L_{\mathbb{N}}$ , the MDL optimal encoding for integers  $z \geq 1$  [26].

Next, we describe how to encode the data in a leaf  $l$ . As we consider both nominal and numeric attributes, we need to define  $L_{nom}(l)$  for nominal and  $L_{num}(l)$  for numeric data.

*Cost of a Nominal Leaf.* To encode the data in a leaf of a nominal attribute, we use Refined MDL [14]. That is, we encode this data minimax optimal, without having to make design choices [5]. In particular, we encode the data using the normalized maximum likelihood (NML) distribution,

$$-\log \left( \frac{\Pr(l \mid \theta^* \in \Theta)}{\sum_{l' \in dom(a_i)^{|l|}} \Pr(l' \mid \theta' \in \Theta)} \right), \quad (3)$$

which encodes data with a code length proportional to how well the best model in the class fits the data at hand, normalized by the sum of maximum likelihoods over all possible data,  $l' \in dom(a_i)^{|l|}$ , each encoded by the best model in the class.

For nominal data, the NML cost for a leaf  $l$  is

$$L_{nom}(l) = \log \left( \sum_{h_1 + \dots + h_k = |l|} \frac{|l|!}{h_1! h_2! \dots h_k!} \right) - |l| \sum_{c \in dom(a_i)} \Pr(a_i = c \mid l) \log \Pr(a_i = c \mid l),$$

where the first term corresponds to the denominator in Eq. (4). Kontkanen & Myllymäki [14] proved the correctness and derived a recursive formula to calculate it in linear time. The second term corresponds to the numerator in Eq. (4), which is instantiated based on the entropy of the leaf.

*Cost of a Numerical Leaf.* For numeric data existing Refined MDL encodings sadly have high computational complexity [14]. Hence, we encode the data in numeric leaves using two-part MDL. In particular, we encode these as point models

assuming Gaussian noise. Note that by this, a split or a regression on an attribute aims to reduce the variance in the leaf. The encoded cost of the data in a numeric leaf, given mean and variance is

$$L_{num}(l \mid \sigma, \mu) = \frac{1}{2\sigma^2 \ln 2} SSE(l, \mu) + \frac{|l|}{2} \log 2\pi\sigma^2 + |l| \log res(a_i) .$$

As we consider empirical data,  $\sigma^2$  and  $\mu$  are estimates over the data in leaf  $l$ . Hence, we can replace  $SSE(l, \mu)$  with  $|l|\sigma^2$  which simplifies the formula to

$$L_{num}(l \mid \sigma, \mu) = \frac{|l|}{2} \left( \frac{1}{\ln 2} + \log 2\pi\sigma^2 \right) + |l| \log res(a_i) .$$

To ensure lossless encoding, we additionally have to encode the model parameters. That is,  $\mu$  and  $\sigma$ . As we consider empirical data, we can safely assume that both lie between the minimum and maximum value of the given attribute. Further, we do not set any prior preference and assume a uniform distribution. Using Kraft's inequality [3], the total cost for a leaf  $l$  then is

$$L_{num}(l) = 2 \log |dom(a_j)| + L_{num}(l \mid \sigma, \mu) .$$

Putting it all together, we now know how to compute  $L(D, M)$ , by which we can formally define the Minimal Coding Forest problem.

**Minimal Coding Forest Problem** *Given a data set  $D$  over a set of attributes  $A = \{a_1, \dots, a_m\}$ , and  $\mathcal{M}$  a valid model class for  $A$ . Find the smallest model  $M \in \mathcal{M}$  such that  $L(D, M)$  is minimal.*

From the fact that both inferring optimal decision trees and structure learning of Bayesian networks—to which our tree-models reduce when considering nominal-only data and split on all values—are NP-hard [18], it follows that the Minimal Coding Forest problem is also NP-hard.

Hence, we resort to heuristics.

## 5 THE CRACK ALGORITHM

Knowing the score  $L(D, M)$  and the problem, we can now introduce the CRACK algorithm, which stands for **classification** and **regression based packing** of data. CRACK is an efficient greedy heuristic for discovering a coding forest  $M$  from given model class  $\mathcal{M}$  with low  $L(D, M)$ . It builds upon the well-known ID3 algorithm [24].

### 5.1 Greedy algorithm

We give the pseudocode of CRACK as Algorithm 1. CRACK starts with an empty model consisting of only trivial trees, i.e. leaf nodes containing all records, per attribute (line 1). The given model class  $\mathcal{M}$  implicitly defines a graph of dependencies between attributes that we are allowed to consider (line 2). That is,  $\mathcal{G}$  is a graph with attributes  $a_i \in A$  as nodes, and with a directed edge from  $a_i$  to  $a_j$  iff there exists a model  $M \in \mathcal{M}$  where attribute  $a_j$  depends on  $a_i$ . To make sure the returned model is valid, we need to maintain a graph representing its dependencies (lines 3–4). We then proceed to iteratively discover that refinement of the current model that maximizes compression. To find the best refinement, we consider every attribute (line 6), and every legal additional split or regression of its corresponding tree (line 10). A refinement is only legal when the dependency is allowed by the model family (line 8), the dependency graph remains acyclic, and we do not split or regress twice on the same attribute (line 9). We keep track of the best found refinement per attribute (lines 11–12), greedily selecting the overall best refinement (line 13), and accepting it only if



**Algorithm 1:** CRACK( $D, \mathcal{M}$ )

---

```

input : data  $D$  over attributes  $A$ , model class  $\mathcal{M}$ 
output: tree model  $M \in \mathcal{M}$  with low  $L(D, M)$ 
1  $T_i \leftarrow \text{TRIVIALTREE}(a_i)$  for all  $a_i \in A$ ;
2  $\mathcal{G} \leftarrow$  dependency graph for  $\mathcal{M}$ ;
3  $V \leftarrow \{v_i \mid i \in A\}$ ,  $E \leftarrow \emptyset$ ;
4  $G \leftarrow (V, E)$ ;
5 while  $L(D, M)$  decreases do
6   for  $a_i \in A$  do
7      $O_i \leftarrow T_i$ ;
8     for  $l \in \text{lhs}(T_i), (i, j) \in \mathcal{G}$  do
9       if  $E \cup (v_i, v_j)$  is acyclic and  $j \notin \text{path}(l)$  then
10         $T'_i \leftarrow \text{REFINELEAF}(T_i, l, j)$ ;
11        if  $L(T'_i) < L(O_i)$  then
12           $O_i \leftarrow T'_i$ ,  $e_i \leftarrow j$ ;
13    $k \leftarrow \arg \min_i \{L(O_i) - L(T_i)\}$ ;
14   if  $L(O_k) < L(T_k)$  then
15      $T_k \leftarrow O_k$ ;
16      $E \leftarrow E \cup (v_k, v_{e_k})$ 
17 return  $M \leftarrow \bigcup_i T_i$ 

```

---

it improves the total encoded length (lines 14–16). If we cannot find any such refinement, we return the best model discovered so far (line 17).

The key subroutine of CRACK is REFINELEAF, in which we discover the optimal refinement of a leaf  $l$  in tree  $T_i$ . That is, it finds the optimal split of  $l$  over all candidate attributes  $a_j$  such that we minimize the encoded length. In case both  $a_i$  and  $a_j$  are numeric, REFINELEAF also considers the best linear and quadratic regression and decides for the variant with the best compression—choosing to split in case of a tie. In the interest of efficiency, we do not allow splitting or regressing multiple times on the same candidate.

## 5.2 Algorithmic complexity

Next we consider the algorithmic complexity of CRACK. In the worst case, we grow a model of full trees, where each candidate splits all leaves on the current height. This means that we have to apply REFINELEAF  $2^m$  times. REFINELEAF is linear in the size of the leaf. For a binary or categorical candidate this follows straightforwardly, as we can compute the NML cost of a split in linear time [14], or even approximate it in sub-linear time [17]. For a numeric leaf we can compute the sum of squared errors in constant time by keeping track of the sum of squares [30]. Therefore, the optimal split can be found in only linear time. This leads to an overall worst case runtime of CRACK of  $O(2^m n)$ . As we only need to store the nodes of the trees, the worst case memory complexity is in  $O(2^m)$ .

Although both complexities look intimidating, CRACK is very fast in practice, taking only up to a few seconds in our experiments. The key reason is that we only consider valid models, and MDL keeps the trees in the models small.

	Binary	Categoric	Numeric	Mixed
CRACK	✓	✓	✓	✓
ORIGO [1]	✓	–	–	–
ERGO [32]	–	–	✓	–
LTR [9]	–	–	✓	–
KTR [2]	–	–	✓	–

Table 1. Data types per multivariate causal inference method.

### 5.3 Causal Inference with CRACK

To compute our causal indicators we have to run CRACK twice on  $D$ . First with model class  $\mathcal{M}_{X \rightarrow Y}$  to obtain  $M_X$  and  $M_{Y|X}$ , and second with  $\mathcal{M}_{Y \rightarrow X}$ , to obtain  $M_Y$  and  $M_{X|Y}$ . With these models we can trivially compute  $\hat{\Delta}_{X \rightarrow Y}$  and  $\hat{\Delta}_{Y \rightarrow X}$ , respectively  $\hat{\delta}_{X \rightarrow Y}$  and  $\hat{\delta}_{Y \rightarrow X}$ , and infer the most likely causal direction. We write  $\text{CRACK}_\delta$ , correspondingly  $\text{CRACK}_\Delta$  to indicate which score we consider.

## 6 RELATED WORK

Causal inference on observational data is a challenging problem as the data at hand was not obtained through controlled randomized experiments. Recently, it has attracted a lot of attention [1, 11, 20, 28]. Most proposals are highly specific in the type of causal dependency and, or type of variables they can consider.

Traditional constrained-based approaches, such as conditional independence tests, require three observed random variables [20, 29], cannot distinguish Markov equivalent causal DAGs [31] and therefore cannot decide between  $X \rightarrow Y$  and  $Y \rightarrow X$ .

Recently, methods have been proposed that can infer the causal direction from only two random variables. Generally, they exploit certain properties of the joint distribution.

Additive Noise Models (ANMs) [28], for example, assume that the effect is a function of the cause and cause-independent additive noise. Causal inference is then done by finding the direction that admits such a model. ANMs exist for univariate real-valued [8, 23, 28, 34] and discrete data [21]. It is unclear how to extend this model for multivariate or mixed type data.

A related approach considers the asymmetry in the joint distribution of *cause* and *effect* for causal inference. The linear trace method (LTR) [9] and the kernelized trace method (KTR) [2] aim to find a structure matrix  $A$  and the covariance matrix  $\Sigma_X$  to express  $Y$  as  $AX$ . Both methods are only applicable to multivariate continuous valued data. In addition, KTR assumes a deterministic, functional and invertible causal relation.

Sgouritsa et al. [27] show that the marginal distribution  $P(\text{cause})$  of the cause does not contain any information about the conditional distribution  $P(\text{effect} \mid \text{cause})$  of the effect. The opposite direction is more likely to contain information. They proposed CURE, which measures this dependency through unsupervised reverse regression for univariate continuous pairs. Liu et al [16] use distance correlation to identify the weakest dependency between univariate pairs of discrete data.

The algorithmic information-theoretic approach views causality in terms of Kolmogorov complexity. The key idea is that if  $X$  causes  $Y$ , the shortest description of the joint distribution  $P(X, Y)$  is given by the separate descriptions of the distributions  $P(X)$  and  $P(Y \mid X)$  [11]. It has also been used in justifying the additive noise model based causal

discovery [12]. However, as Kolmogorov complexity is not computable [15], causal inference using algorithmic information theory requires practical implementations, or notions of independence. For instance, the information-geometric approach [10] defines independence via orthogonality in information space for univariate continuous pairs. Janzing & Schölkopf [11] sketch how comparing marginal distributions, and resource bounded computation could be used to infer causation, but do not give practical instantiations. Vreeken [32] instantiates it with the cumulative entropy to infer the causal direction in continuous univariate and multivariate data. Vreeken and Budhathoki approximate  $K(X)$  and  $K(Y | X)$  through MDL, and propose ORIGO, a decision tree based approach for causal inference on univariate and multivariate binary data [1].

All above univariate methods only consider one data type, but do not combine nominal and numeric data. Table 2 we give an overview of which data types the existing methods for causal inference on multivariate data consider. To the best of our knowledge, CRACK is the first method for causal inference on pairs of univariate or multivariate mixed-type data.

## 7 EXPERIMENTS

In this section, we evaluate CRACK empirically. We implemented CRACK in C++, and provide the source code including the synthetic data generator along with the tested datasets for research purposes.<sup>1</sup> All experiments were executed single-threaded on a MacBook Pro with 2.6 GHz Intel Core i7 processor and 16 GB memory running Mac OS X. All tested data sets could be processed within seconds; over all pairs the longest runtime for CRACK was 3.8 seconds. When applying CRACK to real valued data, we set the resolution parameter globally to that of the attribute with the highest data resolution.

We compare CRACK to DC [16], IGC [10], LTR [9], ORIGO [1] and ERGO [32], using their publicly available implementations.

### 7.1 Synthetic data

We first evaluate  $\text{CRACK}_\delta$  and  $\text{CRACK}_\Delta$  on generated synthetic data with known ground truth. Concretely, we generate univariate or multivariate  $X$  with  $|X|$  and  $Y$  with  $|Y|$  attributes such that  $Y$  depends probabilistically on  $X$ . In the latter, we call this probabilistic relationship *dependency*. As standard setup, we select  $X$  and  $Y$  with the dimensions 5000-by-3 and use *mixed-type* data – meaning that we choose the type per attribute (*binary*, *categorical*, and *real valued*) uniformly at random.

Over all attributes we assume an order such that the attributes of  $X$  are followed by  $Y$ . With the *split probability* we decide whether to refine and attribute by either splitting or regressing on a candidate. We do not allow attributes of  $X$  to depend on  $Y$ . Further, we control the probability for attributes of  $Y$  to depend on  $X$  (*dependency*). The stronger the dependency, the higher the probability that the induced ground truth  $X \rightarrow Y$  holds.

Using the trees, we generate the data randomly per leaf. For binary data we choose the percentage of ones uniformly at random. Categorical data is restricted to three to four values, the frequencies for which we again generate uniformly at random. Real valued data is generated with a normal distribution by choosing a random mean and standard deviation. We restrict ourselves to the domain  $[0, 5]$ , as larger values lead to overly obvious dependencies.

As it has been shown in ORIGO [1] and discussed in Section 3 the  $\hat{\Delta}_{X \rightarrow Y}$  score is more balanced with regard to multivariate data. We therefore use  $\text{CRACK}_\Delta$  for the performance tests and later compare both scores.

<sup>1</sup><http://eda.mmci.uni-saarland.de/crack/>

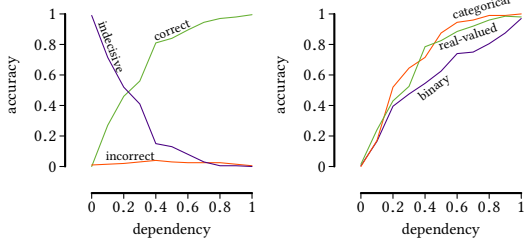


Fig. 3. (left) Fraction of correct, incorrect and indecisive decisions. (right) Correct decisions for binary, categorical and real valued data.

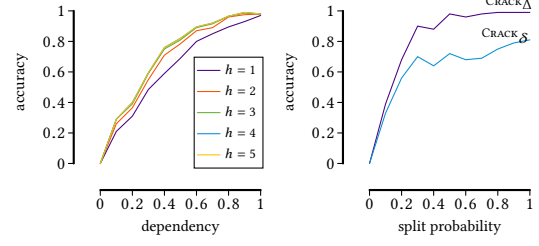


Fig. 4. [Higher is better] Accuracy of CRACK on synthetic data of different tree heights (left) and with different split probabilities (right).

*Performance.* First, we investigate the performance of  $\text{CRACK}_\Delta$  on data with varying dependency. To this end, we generate for each dependency level (0.0, 0.1, ..., 1.0) 200 random data sets fixing the split probability to 1.0. Further, we allow the trees to have maximum height. We report the fraction of correct inferences (*accuracy*), the fraction of incorrect inferences and the fraction of indecisive inferences and plot them in Figure 3. We see that at a dependency level of 0 we correctly infer there is no causal direction. The fraction of correct inferences rises steeply if the dependency increases while we almost make no incorrect inferences. After a dependency of 0.6, the precision is over 90%.

Looking more closely at the performance per data type, we observe in Figure 3 that the accuracy for categorical and real valued data is very similar, exceeding 90% accuracy at 0.65 dependency. For binary data, we observe a slower increase in the accuracy, which is due to the lower diversity (smaller domain) of binary data compared to categorical and real valued data.

*Robustness.* To evaluate the robustness of CRACK we perform two tests. First, we restrict the height of the trees, but keep the other parameters as before. Figure 4 shows that CRACK works already well when the generating trees have only one split. For two splits CRACK performs comparable to when we do not bound the number of splits.

Second, we compare  $\text{CRACK}_\Delta$  and  $\text{CRACK}_\delta$  on real valued data of 2 000 rows, with dependency 1, and varying the split probability. We report the average accuracy over 100 trials in Figure 4. We see that when the generating model fits our score,  $\text{CRACK}_\Delta$  outperforms  $\text{CRACK}_\delta$ , reaching an accuracy of 78% at a split probability of 0.3.

*Dimensionality.* The next tests are designed to evaluate both  $\text{CRACK}_\delta$  and  $\text{CRACK}_\Delta$  performs pairs of varying dimensionality. We fix the split probability and the dependency to 1.0 and set the number of data points per attribute to 5 000. We first test symmetric pairs of real-valued attributes, where  $|X| = |Y|$ , varying the dimensions (1 to 10). We plot the results in Figure 5. We see that both scores work well, with  $\text{CRACK}_\Delta$  leading at a small margin. More specifically, for univariate pairs  $\text{CRACK}_\Delta$  has about 0.8 accuracy, whereas if 3 or more dimensions are considered it approaches perfect accuracy.

Next, we consider pairs of asymmetric cardinality, keeping all other parameters the same. In particular, we fix  $|X| = 5$  and vary  $|Y|$  as before. To avoid any bias towards the dimensionality of the effect, we perform both 100 runs with the consequent having the smaller dimension and 100 runs with the cause having the smaller dimension. We plot the results in Figure 5. As expected, we see that  $\text{CRACK}_\Delta$  does not suffer from the asymmetry, whereas  $\text{CRACK}_\delta$  performs worse the larger the asymmetry.

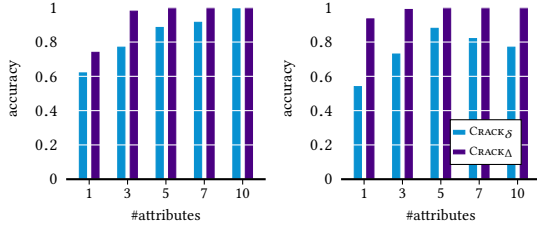


Fig. 5. [Higher is better] Accuracy of CRACK on pairs of (left) symmetric cardinality,  $|X| = |Y|$ , and (right) asymmetric cardinality, randomly fixing  $|X| = 5$  or  $|Y| = 5$ .

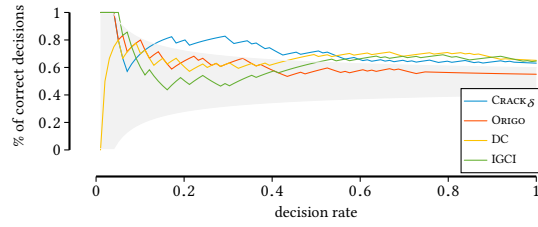


Fig. 6. [Higher is better] Accuracy per decision rate for CRACK $_{\delta}$ , ORIGO, DC, and IGCI on 99 univariate pairs from the Tübingen benchmark dataset.

## 7.2 Real world data

To evaluate CRACK $_{\delta}$  and CRACK $_{\Delta}$  on real world data, we consider first univariate and second multivariate pairs.

*Univariate pairs.* We first evaluate CRACK on all 99 univariate pairs available in version 1.0 of the Tübingen database.<sup>2</sup> We compare CRACK to IGCI [10], DC [16], and ORIGO [1]. For each method we sort the pairs descending according to their decision strength. If an algorithm did not decide for a causal direction, we weighted the results as 0.5. To apply ORIGO, we discretized the data with IPD [19] as proposed by the authors. For DC we discretized the data as described in the paper.

We plot the corresponding decision rate—the percentage of correct decisions over the top- $k$  pairs with highest difference in scores  $X \rightarrow Y$  and  $Y \rightarrow X$ —together with the 95% confidence interval for a random coin flip in Figure 6. Many of these pairs have rather unbalanced domain sizes. As discussed in Section 3.5 we expect CRACK $_{\delta}$  to perform better on these data, and indeed found this to be the case. To avoid clutter, we only show the curve for CRACK $_{\delta}$ .

We see that CRACK $_{\delta}$  performs rather favourably compared to its competitors. At one third of all pairs its accuracy is 83%, and overall its performance is significantly better than random for 90% of decisions. Comparing between the methods, we find that CRACK $_{\delta}$  beats IGCI significantly over 29.3%, and beats DC significantly over 21.2% of all pairs, with regard to the 95% confidence interval over the decision rate of CRACK. Over all pairs IGCI or DC perform on par with CRACK $_{\delta}$ , but never outperform us significantly.

*Multivariate pairs.* Second, we evaluate on twelve multivariate cause-effect pairs from several sources. The first five (*Climate forecast*, *Ozone*, *Car efficiency*, *Radiation*, *Symptoms* and *Brightness*) belong to the Tübingen cause-effect pairs. *Chemnitz* and *Precipitation* were used before by Janzing et al. [9]. *Haberman* is a data set on medical case studies describing the survival of patients who had undergone surgery for breast cancer between 1958 and 1970 [6].  $X$  consists of the age of the patient at time of operation, the patient’s year of operation and the number of positive axillary nodes detected.  $Y$  is the survival status, which is binary and divided into longer or at most five years ( $X \rightarrow Y$ ). The *Iris* data set contains data about three types of the Iris plant ( $Y$ ) and four features dependent on which the type can be determined [4]. Last, we extract two cause-effect data sets from the Mammals data set [7], which consists of both climate data and presence records of 121 mammal species over 2183 areas of  $50 \times 50$ km in Europe. We assume that elevation, precipitation, average temperature and the annual temperature range ( $X$ ) cause the presence of a mammal and not contrarily. For *Canis* we selected two animals of the canis (wolf) family, and for *Lepus* three animals from the

<sup>2</sup><https://webdav.tuebingen.mpg.de/cause-effect/>

Dataset	$m$	$k$	$l$	Decisions per method				
				LTR	ERGO	ORIGO	CRACK <sub><math>\delta</math></sub>	CRACK <sub><math>\Delta</math></sub>
Climate	10 226	4	4	✓	✓	–	✓	–
Ozone	989	1	3	(n/a)	✓	✓	✓	✓
Car	392	3	2	–	✓	✓	–	✓
Radiation	72	16	16	–	–	–	✓	✓
Symptoms	120	6	2	✓	✓	–	✓	✓
Brightness	1 000	9	1	(n/a)	(n/a)	–	✓	–
Chemnitz	1 440	3	7	✓	✓	✓	–	✓
Precip.	4 748	3	12	✓	–	–	–	✓
Haberman	306	2	2	✓	✓	–	–	✓
Iris flower	150	4	1	(n/a)	(n/a)	–	✓	✓
Canis	2 183	6	2	(n/a)	(n/a)	✓	✓	✓
Lepus	2 183	6	3	(n/a)	(n/a)	✓	✓	✓
<b>Accuracy</b>				0.42	0.50	0.42	0.67	0.83

Table 2. Comparison of LTR, ERGO, ORIGO, CRACK <sub>$\delta$</sub>  and CRACK <sub>$\Delta$</sub>  on eleven multivariate data sets. We write (n/a) whenever a method is not applicable on the pair.

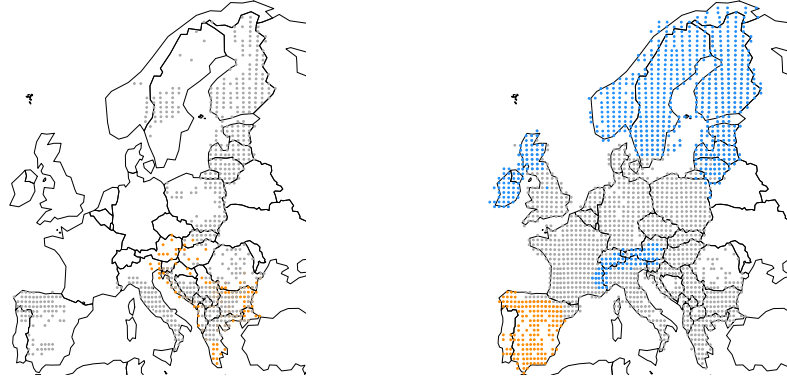


Fig. 7. Presence records in Europe of (left) Grey Wolf (grey) and Golden Jackal (gold), and (right) Mountain hare (gold), European hare (grey) and Granada hare (blue).

lepus (hare) family. We plotted presence of these animals in Figure 7. Both mammals data sets, as well as the *Symptoms* data set contain only binary attributes in  $Y$ . To avoid bias due to the domain sizes of the real valued features (see Sec. 3.5), we normalized the real valued features between zero and one.

We compare CRACK to LTR [9], ERGO [32], and ORIGO [1], applying these where applicable. We give the characteristics of the data sets, as well as the results in Table 2.

Overall, CRACK <sub>$\Delta$</sub>  performs best with an accuracy of 83% and CRACK <sub>$\delta$</sub>  performs well on the near-symmetric pairs having an overall accuracy of 67%. LTR and ERGO perform similar on pure numeric pairs but can not deal with those pairs containing binary or categorical data.

## 8 DISCUSSION

The experiments show that CRACK works very well in practice. On synthetic data  $\text{CRACK}_\delta$  and  $\text{CRACK}_\Delta$  both identify the ground truth with high accuracy, even on data with relatively weak and few dependencies. Evaluation on univariate real-valued benchmark pairs shows that  $\text{CRACK}_\delta$  outperforms the state of the art significantly over a large interval of decisions. Over 12 mixed-type multivariate pairs,  $\text{CRACK}_\Delta$  recovers the ground truth with an accuracy of 83%.

The performance of CRACK is particularly impressive if we take into account its simplicity. In the interest of computational efficiency, we only consider binary splits on single attribute values, and are restricted to using an attribute only once per path in a tree. At the cost of extra computation, multi-way and interval splits will likely improve performance. Similarly, it will be interesting to see if e.g. Gaussian-process or kernelized regression will improve inference accuracy.

Ideally, we would use Refined MDL to approximate Kolmogorov complexity. We are, however, not aware of an efficiently computable score for coding forests. We therefore constructed a two-part MDL encoding, which involves choices—alternate choices may be more efficient, and may lead to better models. For example, it will be interesting to see whether the ideas of Wallace [33] for decision tree encoding can be used to improve CRACK, as well as to explore efficient ways to compute the NML cost for numeric leaves.

It will be interesting to consider CRACK for causal structure learning. That is, applying CRACK to a data set without knowing  $X$  and  $Y$  and mine cause effect pairs based on the strength of the edges in the DAG. Another interesting direction to explore is that of time series data, which would require to extend our current framework with temporal dependencies.

## 9 CONCLUSION

We considered the problem of inferring the causal direction from the joint distribution of two univariate or multivariate random variables  $X$  and  $Y$  consisting of single, or mixed type data. To infer the causal direction we took an information theoretic approach identifying the most likely causal direction as the one with the most succinct code length. We proposed a practical encoding scheme based on MDL to describe nominal and numeric data and model dependencies between  $X$  and  $Y$  using classification and regression trees. Further, we introduced CRACK, a fast greedy heuristic to infer the causal direction for mixed type data.

Experiments show that CRACK reliably infers the correct causal direction with high confidence. On multivariate real world data, we outperform the state of the art and on univariate benchmark data CRACK performs at least as well as univariate single type methods. In future work, we are curious to investigate in causal discovery, that is, to directly identify cause effect pairs from a data set in which  $X$  and  $Y$  are not known.

## ACKNOWLEDGMENTS

The authors wish to thank Kailash Budhathoki for insightful discussions. Alexander Marx is supported by the International Max Planck Research School for Computer Science (IMPRS-CS). Both authors are supported by the Cluster of Excellence “Multimodal Computing and Interaction” within the Excellence Initiative of the German Federal Government.

## REFERENCES

- [1] Kailash Budhathoki and Jilles Vreeken. 2016. Causal Inference by Compression. In *Proceedings of the 16th IEEE International Conference on Data Mining (ICDM), Barcelona, Spain*. IEEE, 41–50.

- [2] Z. Chen, K. Zhang, and L. Chan. 2013. Nonlinear Causal Discovery for High Dimensional Data: A Kernelized Trace Method. In *Proceedings of the 13th IEEE International Conference on Data Mining (ICDM)*, Dallas, TX. IEEE, 1003–1008.
- [3] Thomas M. Cover and Joy A. Thomas. 2006. *Elements of Information Theory*. Wiley-Interscience New York.
- [4] Ronald A Fisher. 1936. The use of multiple measurements in taxonomic problems. *Annals of eugenics* 7, 2 (1936), 179–188.
- [5] Peter Grünwald. 2007. *The Minimum Description Length Principle*. MIT Press.
- [6] Shelby J Haberman. 1976. Generalized residuals for log-linear models. In *Proceedings of the 9th international biometrics conference*. 104–122.
- [7] H Heikinheimo, M Fortelius, J Eronen, and H Mannila. 2007. Biogeography of European land mammals shows environmentally distinct and spatially coherent clusters. *Journal of Biogeography* 34 (2007), 1053–1064. Issue 6.
- [8] PO. Hoyer, D. Janzing, JM. Mooij, J. Peters, and B. Schölkopf. 2009. Nonlinear causal discovery with additive noise models. 689–696.
- [9] D. Janzing, P. Hoyer, and B. Schölkopf. 2010. Telling cause from effect based on high-dimensional observations. In *Proceedings of the 27th International Conference on Machine Learning (ICML)*, Haifa, Israel. JMLR, 479–486.
- [10] Dominik Janzing, Joris Mooij, Kun Zhang, Jan Lemeire, Jakob Zscheischler, Povilas Daniušis, Bastian Steudel, and Bernhard Schölkopf. 2012. Information-geometric approach to inferring causal directions. *Artificial Intelligence* 182-183 (2012), 1–31.
- [11] D. Janzing and B. Schölkopf. 2010. Causal Inference Using the Algorithmic Markov Condition. *IEEE Transactions on Information Technology* 56, 10 (2010), 5168–5194.
- [12] D. Janzing and B. Steudel. 2010. Justifying Additive Noise Model-Based Causal Discovery via Algorithmic Information Theory. *Open Systems and Information Dynamics* 17, 2 (2010), 189–212.
- [13] A.N. Kolmogorov. 1965. Three Approaches to the Quantitative Definition of Information. *Problemy Peredachi Informatsii* 1, 1 (1965), 3–11.
- [14] P. Kontkanen and P. Myllymäki. 2007. MDL histogram density estimation. In *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics (AISTATS)*, San Juan, Puerto Rico. JMLR, 219–226.
- [15] M. Li and P. Vitányi. 1993. *An Introduction to Kolmogorov Complexity and its Applications*. Springer.
- [16] Furui Liu and Laiwan Chan. 2016. Causal Inference on Discrete Data via Estimating Distance Correlations. *Neural Computation* 28, 5 (2016), 801–814.
- [17] Tommi Mononen and Petri Myllymäki. 2008. Computing the Multinomial Stochastic Complexity in Sub-Linear Time. In *Proceedings of the 4th European Workshop on Probabilistic Graphical Models*. 209–216.
- [18] Kolluru Venkata Sreerama Murthy. 1997. *On Growing Better Decision Trees from Data*. PhD Thesis. Johns Hopkins University, Baltimore.
- [19] Hoang-Vu Nguyen, Emmanuel Müller, Jilles Vreeken, and Klemens Böhm. 2014. Unsupervised Interaction-Preserving Discretization of Multivariate Data. *Data Mining and Knowledge Discovery* 28, 5 (2014), 1366–1397.
- [20] Judea Pearl. 2009. *Causality: Models, Reasoning and Inference* (2nd ed.). Cambridge University Press, New York, NY, USA.
- [21] J. Peters, D. Janzing, and B. Schölkopf. 2010. Identifying Cause and Effect on Discrete Data using Additive Noise Models. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*. JMLR, 597–604.
- [22] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. 2011. Causal Inference on Discrete Data using Additive Noise Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33, 12 (2011), 2436–2450.
- [23] J. Peters, JM. Mooij, D. Janzing, and B. Schölkopf. 2014. Causal Discovery with Continuous Additive Noise Models. *Journal of Machine Learning Research* 15 (2014), 2009–2053.
- [24] J Ross Quinlan. 1986. Induction of decision trees. *Machine Learning* 1, 1 (1986), 81–106.
- [25] Jorma Rissanen. 1978. Modeling by shortest data description. *Automatica* 14, 1 (1978), 465–471.
- [26] Jorma Rissanen. 1983. A Universal Prior for Integers and Estimation by Minimum Description Length. *The Annals of Statistics* 11, 2 (1983), 416–431.
- [27] Eleni Sgouritsa, Dominik Janzing, Philipp Hennig, and Bernhard Schoelkopf. 2015. Inference of Cause and Effect with Unsupervised Inverse Regression. *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)* 38 (2015), 847–855.
- [28] Shohei Shimizu, Patrik O. Hoyer, Aapo Hyvärinen, and Antti Kerminen. 2006. A Linear Non-Gaussian Acyclic Model for Causal Discovery. *Journal of Machine Learning Research* 7 (2006), 2003–2030.
- [29] P. Spirtes, C. Glymour, and R. Scheines. 2000. *Causation, Prediction, and Search*. MIT press.
- [30] Evimaria Terzi. 2006. *Problems and Algorithms for Sequence Segmentations*. PhD thesis. University of Helsinki.
- [31] Thomas Verma and Judea Pearl. 1991. Equivalence and Synthesis of Causal Models. In *Proceedings of the 6th International Conference on Uncertainty in Artificial Intelligence (UAI)*. North-Holland, 255–270.
- [32] Jilles Vreeken. 2015. Causal Inference by Direction of Information. In *Proceedings of the SIAM International Conference on Data Mining (SDM)*, Vancouver, Canada. SIAM, 909–917.
- [33] C.S. Wallace and J.D. Patrick. 1993. Coding Decision Trees. *Machine Learning* 11, 1 (1993), 7–22. DOI: <http://dx.doi.org/10.1023/A:1022646101185>
- [34] Kun Zhang and Aapo Hyvärinen. 2009. On the Identifiability of the Post-nonlinear Causal Model. In *Proceedings of the 25th International Conference on Uncertainty in Artificial Intelligence (UAI)*. AUAU Press, 647–655.